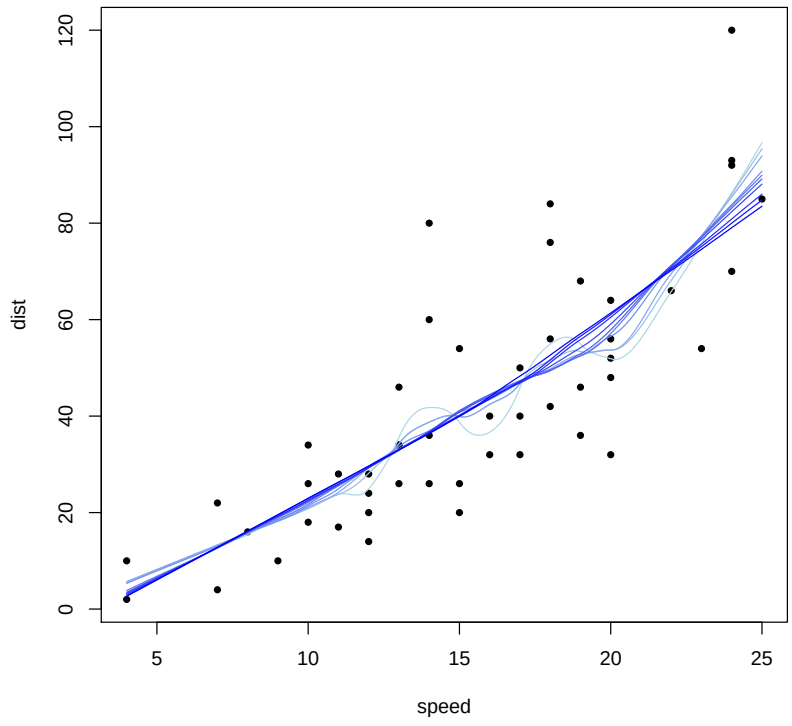
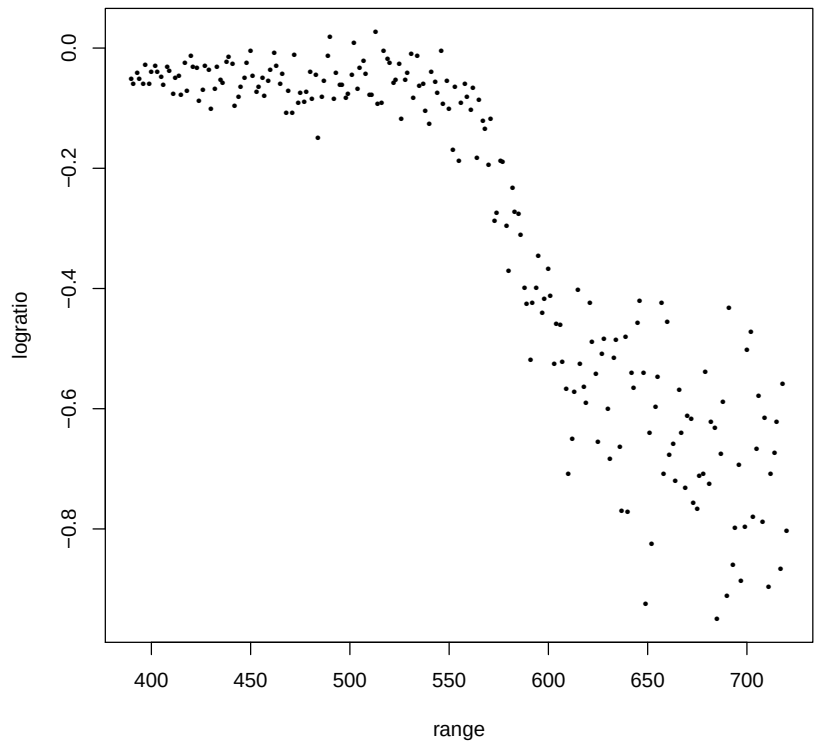
# Modélisation Non Linéaire: Introduction




Yannig Goude

EDF R&D -dpt OSIRIS- Clamart

# *Exemple: lidar data*



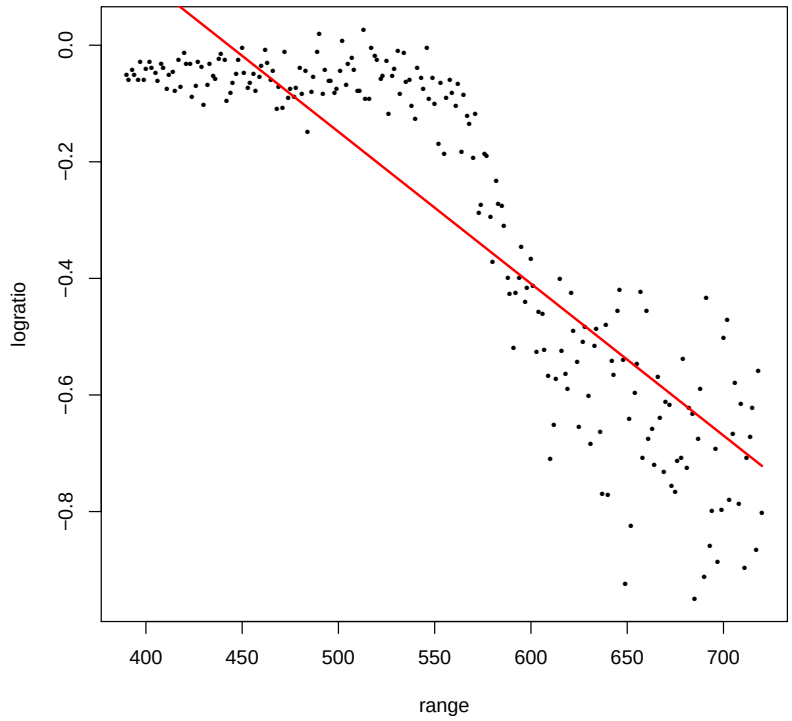
Wikipédia: *light detection and ranging, est une technologie de télédétection ou de mesure optique basée sur l'analyse des propriétés d'une lumière laser renvoyée vers son émetteur.*

Technique utilisée pour détecter des composants chimiques dans l'atmosphère (polluant notamment).

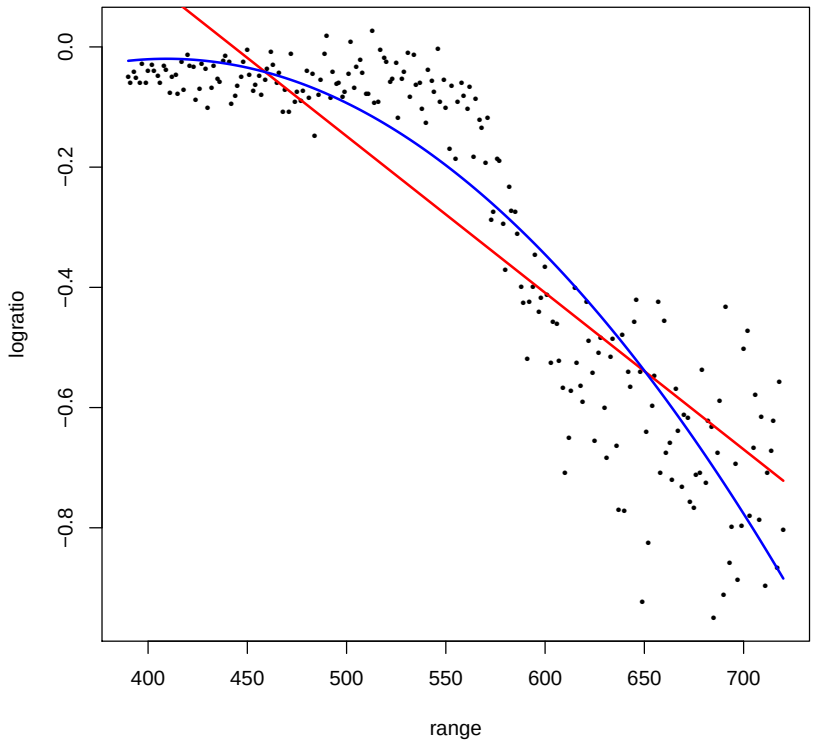
- la variable *range* est la distance parcourue avant que la lumière revienne au laser
- la variable *logratio* est le log du rapport de lumière reçue par 2 sources laser. Une source à une fréquence de résonance correspondant au composant d'intérêt -ici le mercure-, l'autre non.

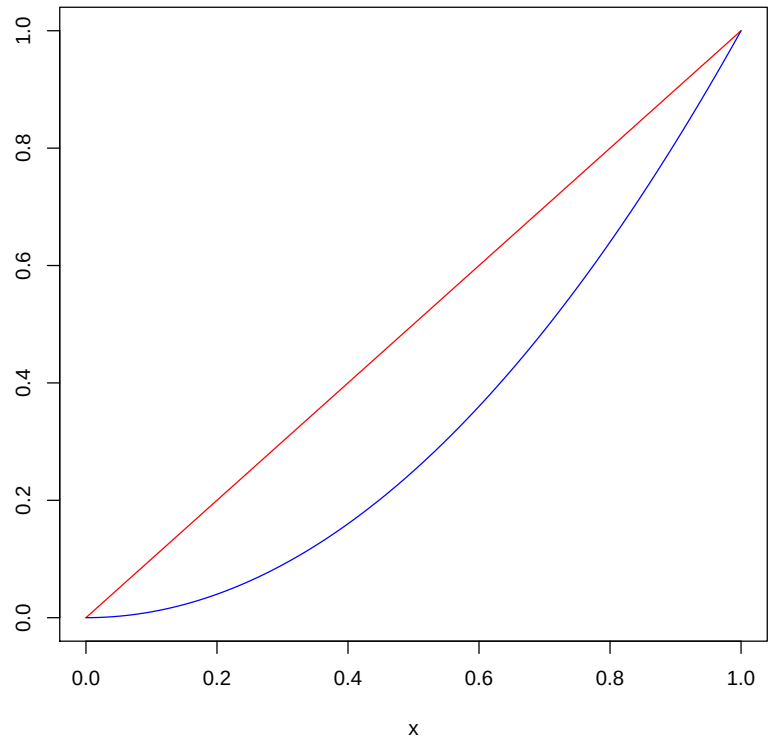
la concentration d'un composant à une distance donné est donc proportionnelle à la pente de la courbe estimée.

*Exemple1: lidar data set, régression linéaire sur* $\mathbf{1}, x$

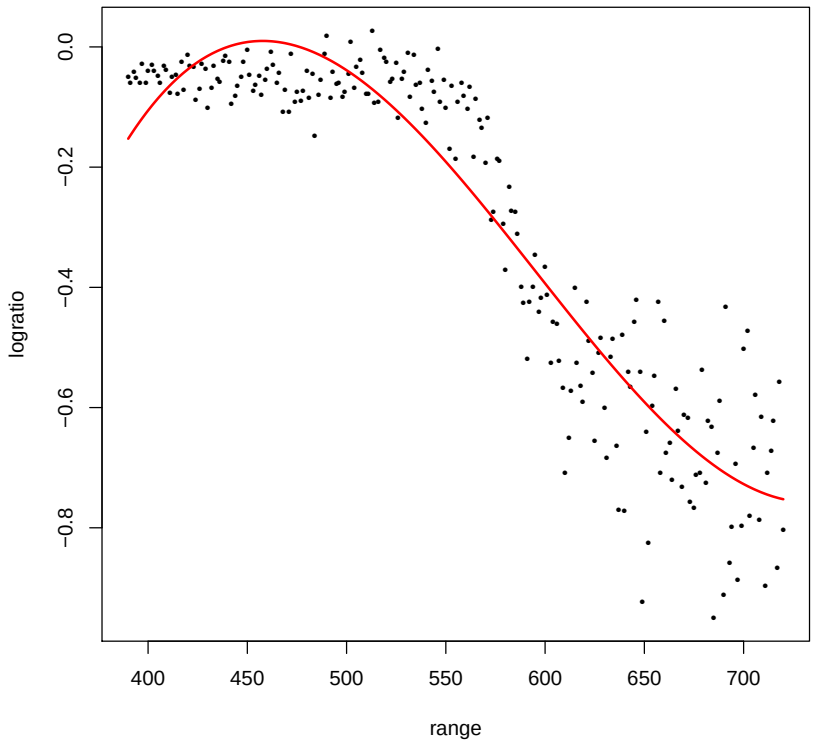


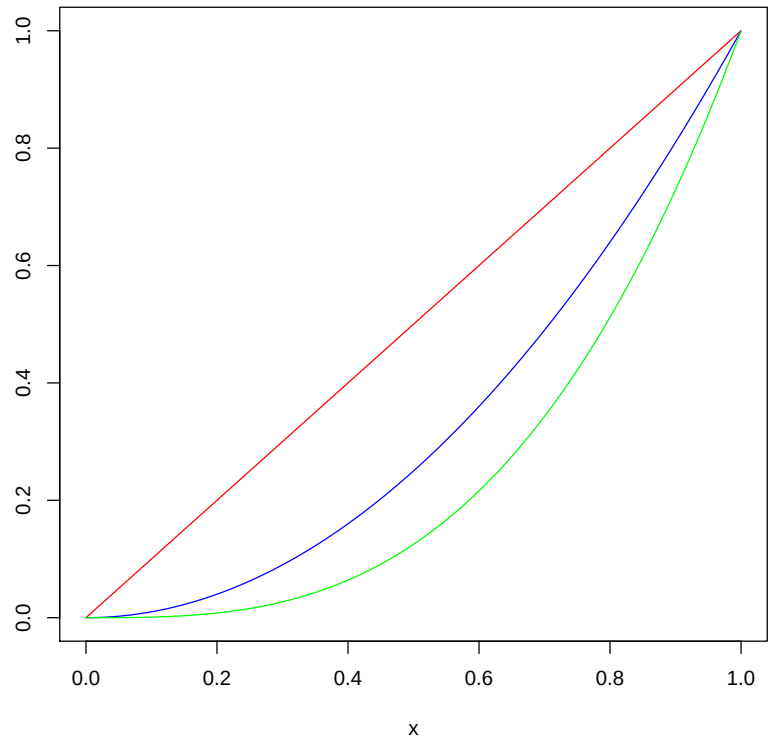
# *Exemple1: modélisation linéaire sur:* $\mathbf{1}$, $x$ *et* $x^2$



*Exemple1: trans. polynômiale, reg. linéaire sur:* $\mathbf{1}$, $x$, $x^2$ *et* $x^3$



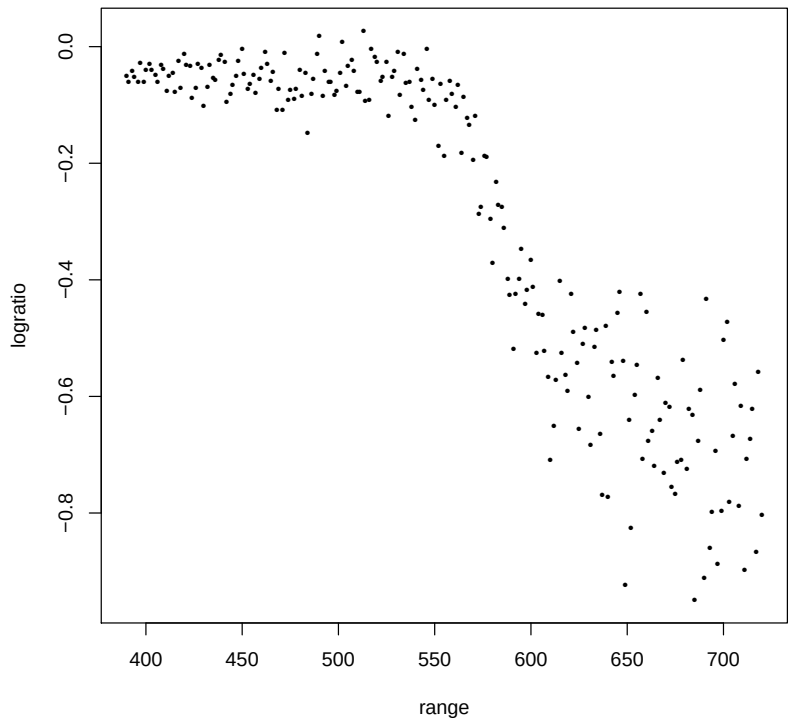
*Kernel Regression*

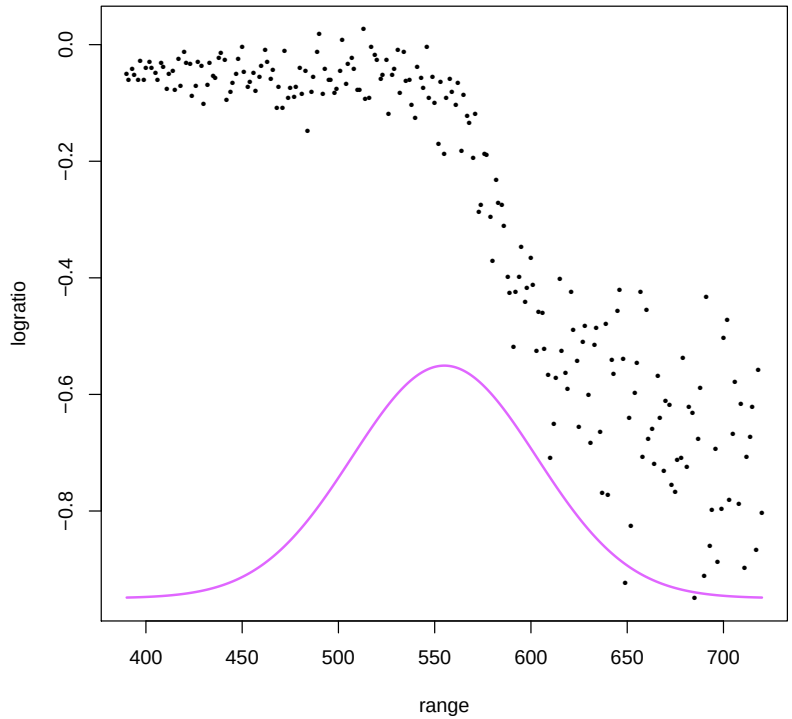
$$\widehat{s}(x) = \sum_{i=1}^{n} w_i(x) y_i$$

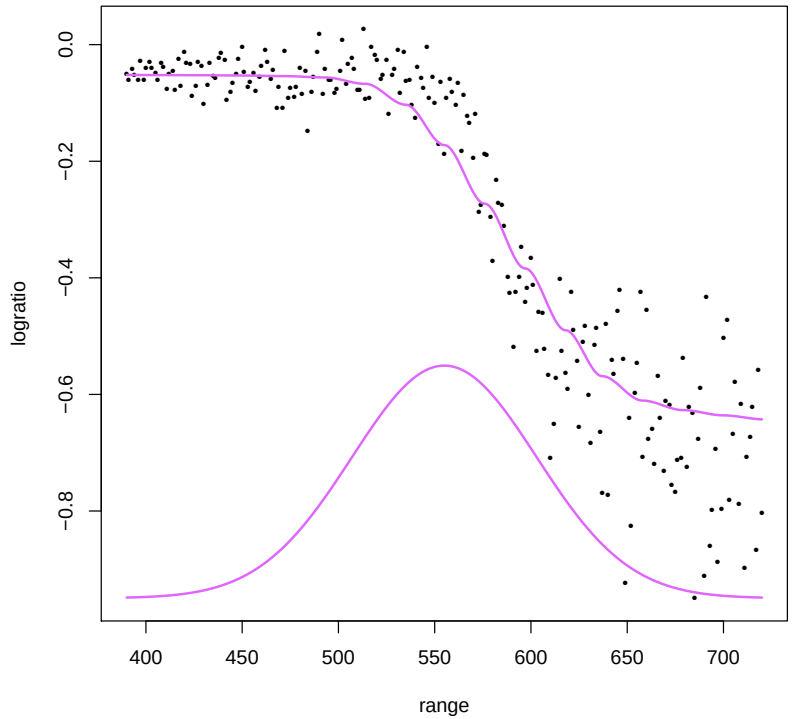
Avec:

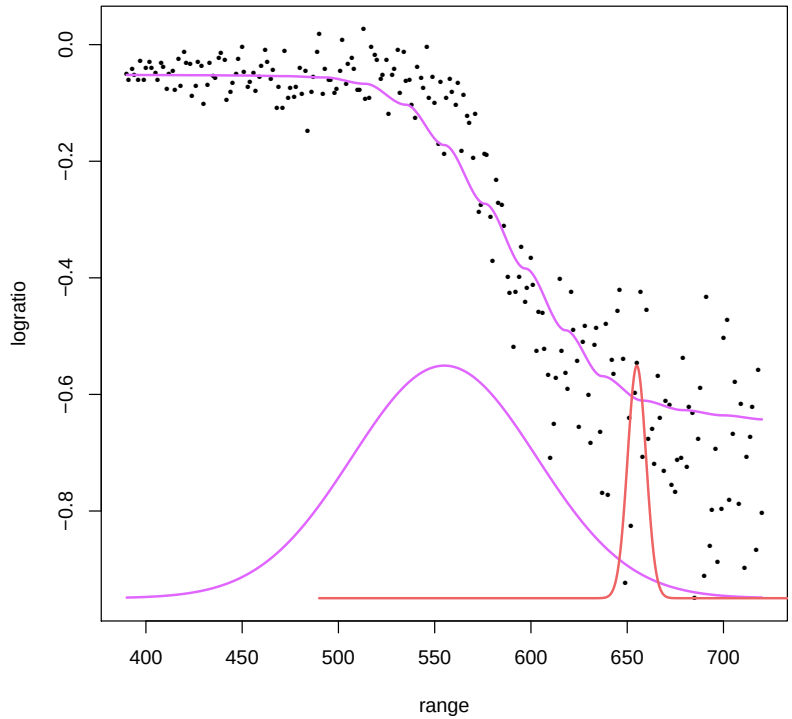
$$w_i(x) \sim \exp(-(x - x_i)^2/h)$$

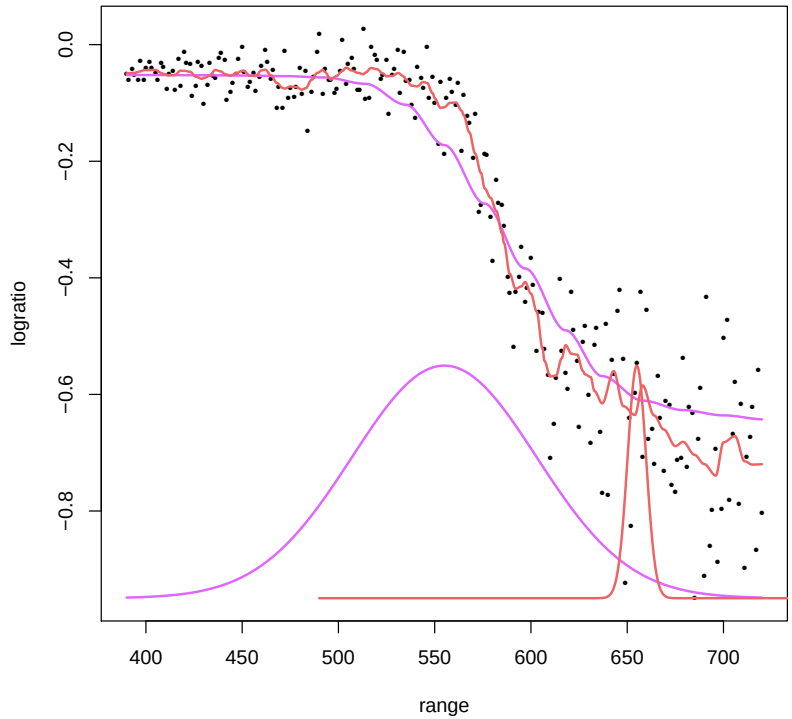
- $\widehat{s}(x)$ est une moyenne pondérée des $y$
- plus une observation est proche de $x$ plus son poids est important
- $h$: la "fenêtre" détermine ou le degré de lissage de l'estimateur, sa *régularité*

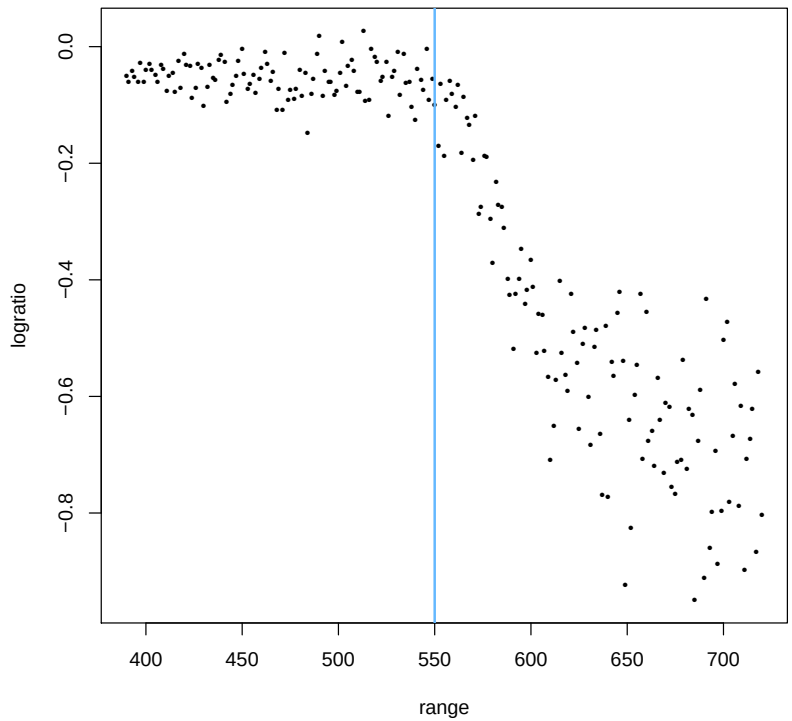

logratio
0.0
−0.2
−0.4
−0.6
−0.8
400
450
500
550
600
650
700
range

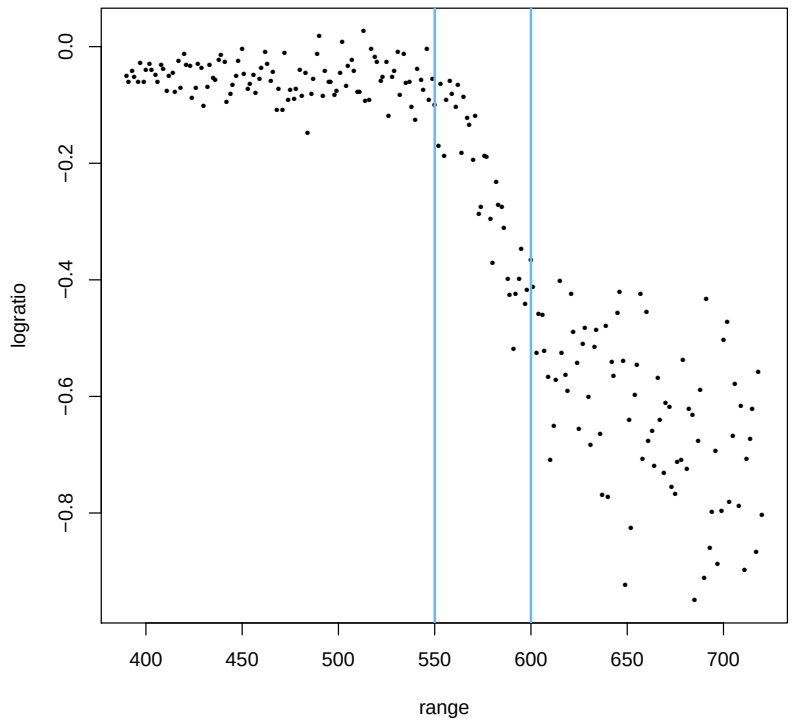

logratio
0.0
−0.2
−0.4
−0.6
−0.8
400
450
500
550
600
650
700
range


0.0
−0.2
−0.4
−0.6
−0.8
logratio
400
450
500
550
600
650
700
range

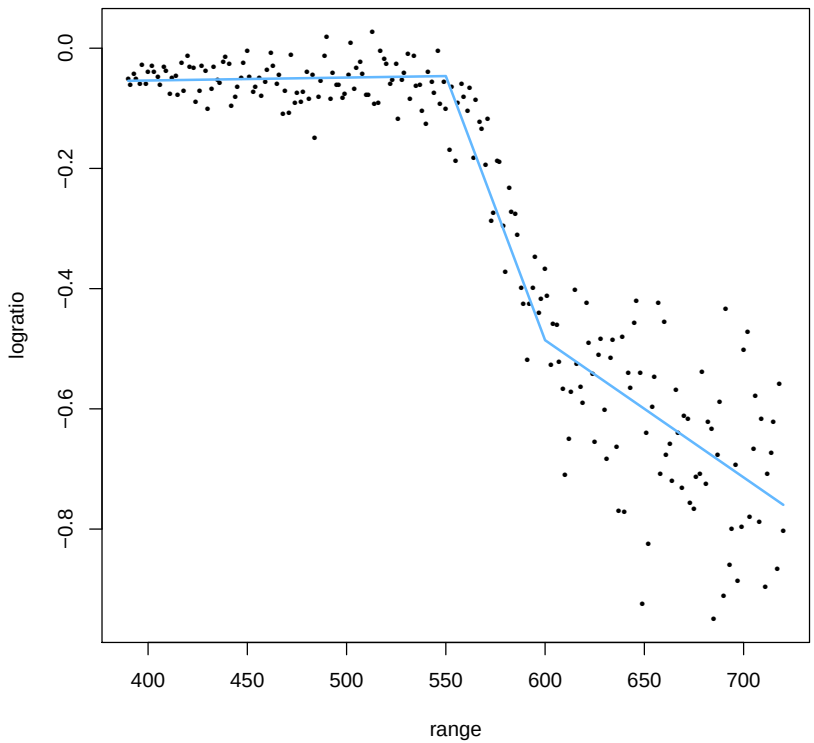

logratio
0.0
−0.2
−0.4
−0.6
−0.8
400
450
500
550
600
650
700
range

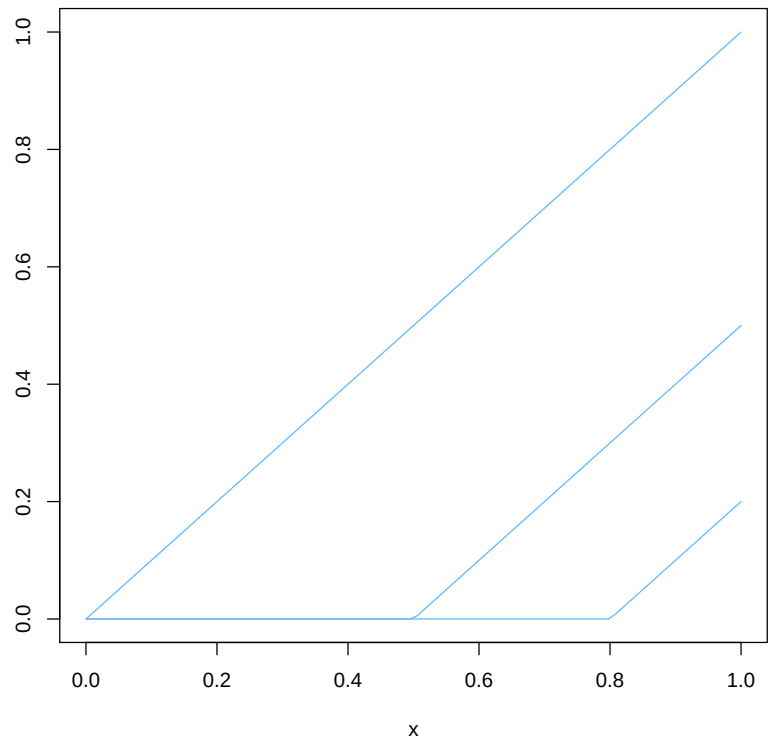

logratio
0.0
−0.2
−0.4
−0.6
−0.8
400
450
500
550
600
650
700
range


logratio
0.0
−0.2
−0.4
−0.6
−0.8
400
450
500
550
600
650
700
range


logratio
0.0
−0.2
−0.4
−0.6
−0.8
400
450
500
550
600
650
700
range

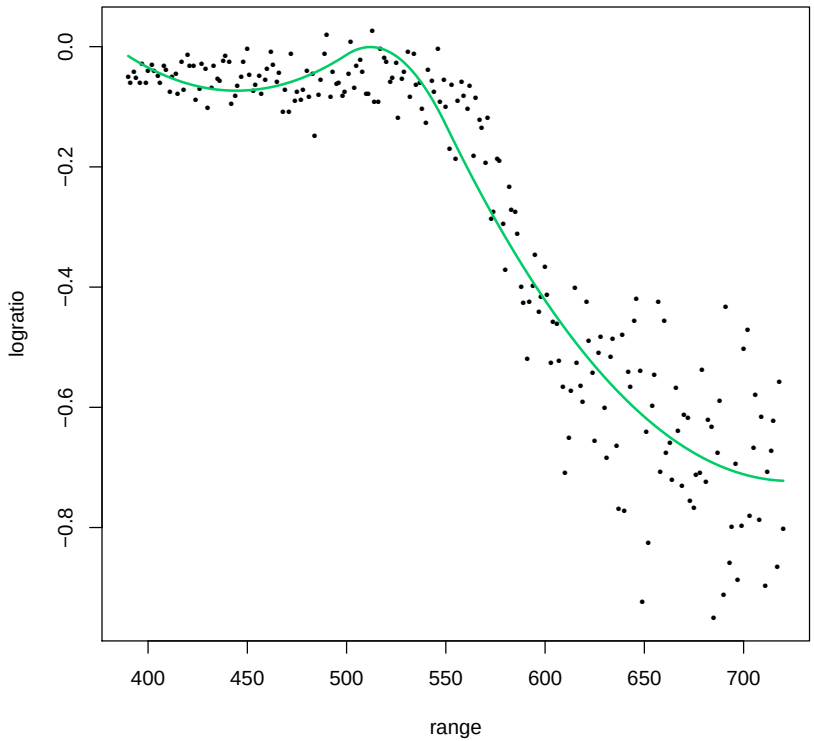
*Exemple1: régression linéaire sur* $\mathbf{1}$,$x$, $(x-550)_+$, $(x-600)_+$

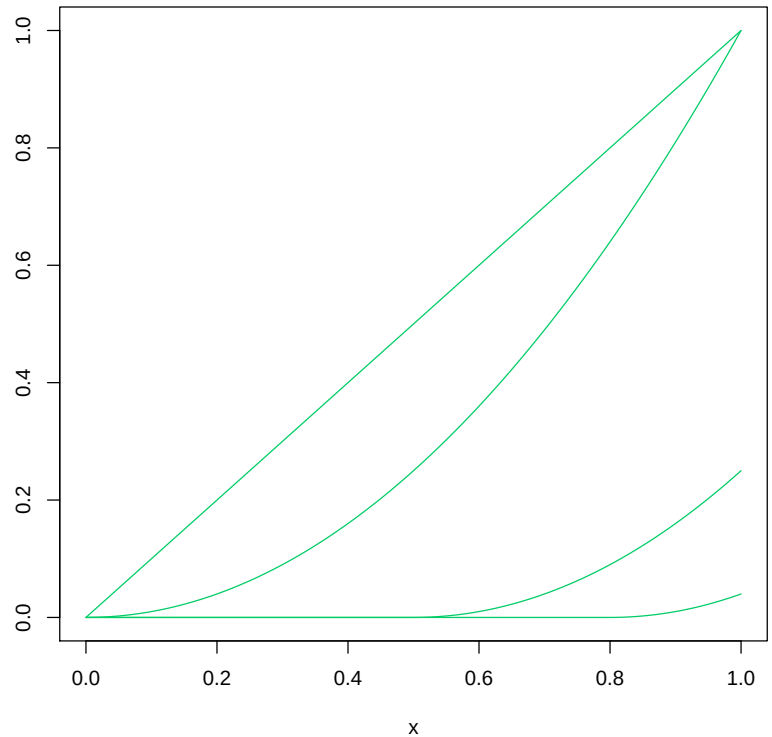



550 et 600 sont appelés des *noeuds*, et les fonctions $(x-k)_+$ sont appelées *splines linéaires*. L'ensemble des fonctions $(x-k)_+$ est une base de spline linéaire.

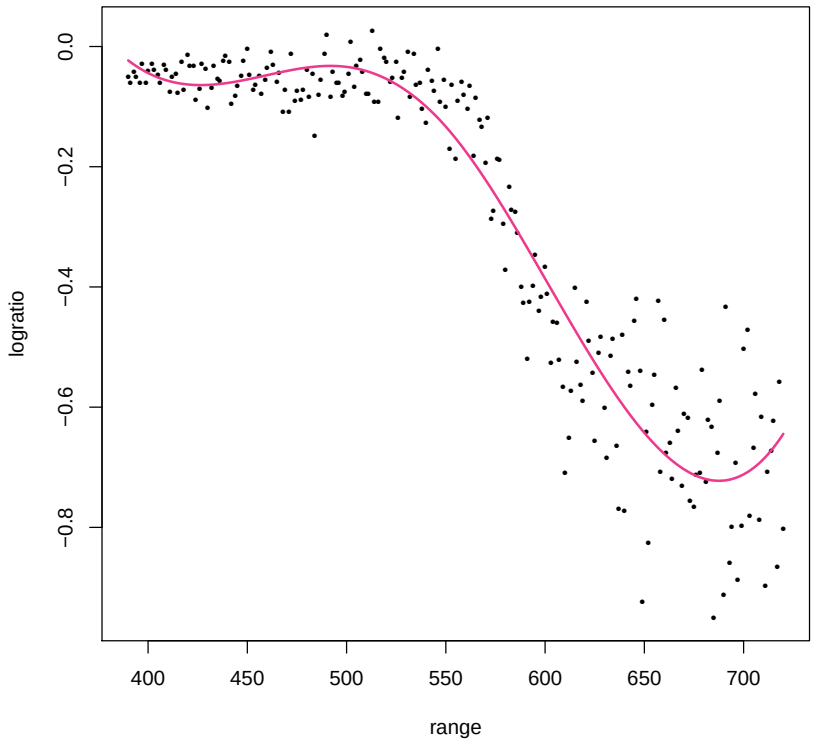
*Exemple1: régression linéaire sur* $\mathbf{1}$,$x$, $x^2$, $(x-550)_+^2$, $(x-600)_+^2$

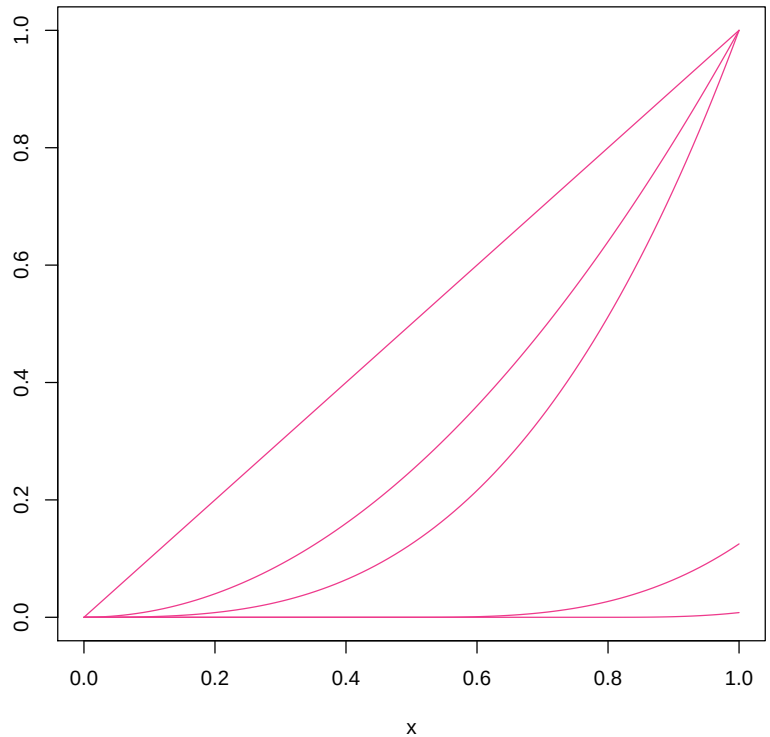



les fonctions $(x-k)_+^2$ sont appelées *splines quadratiques*

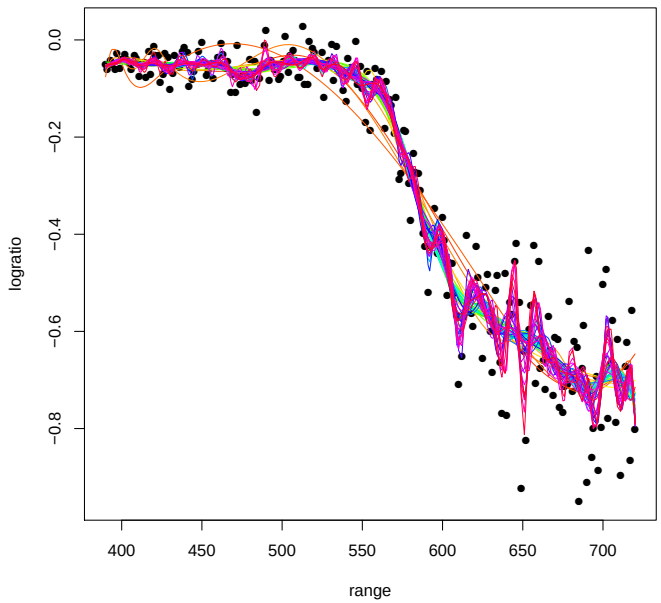
*Exemple1: régression linéaire sur* $\mathbf{1}$,$x$, $x^2$, $x^3$, $(x-550)^3_+$, $(x-600)^3_+$




les fonctions $(x-k)^3_+$ sont appelées *splines cubiques*

Pb: choix du nombre et de la position des "noeuds", du degré des splines




2 alternatives:

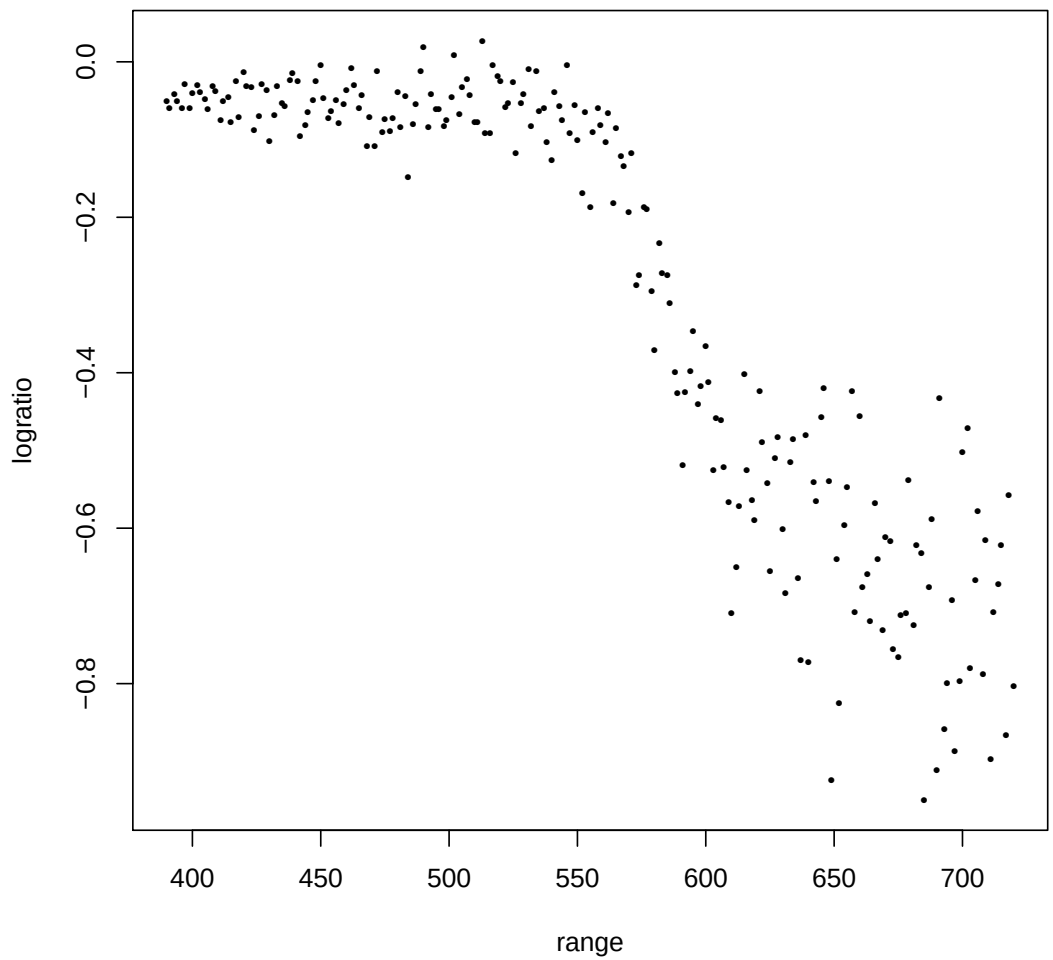
- Sélection de modèle
- Méthode de pénalisation

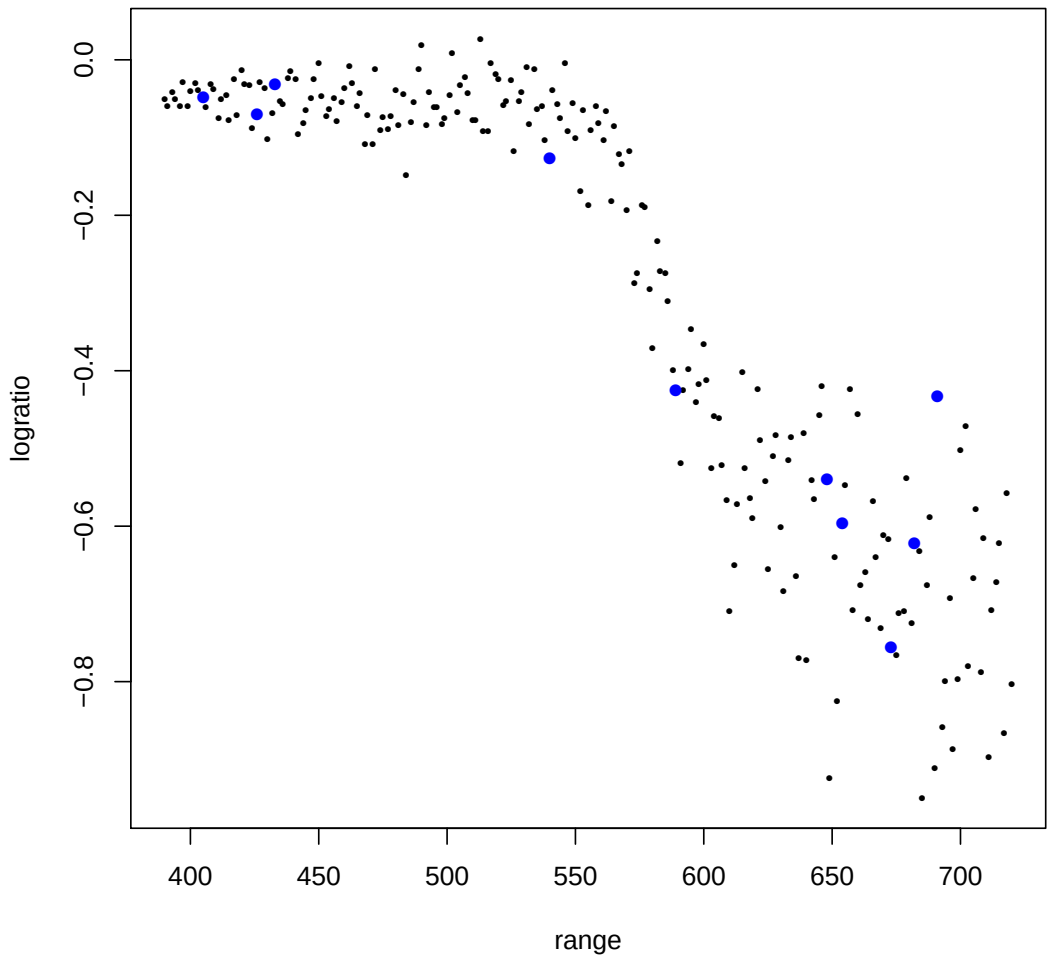
*Sélection de modèle*

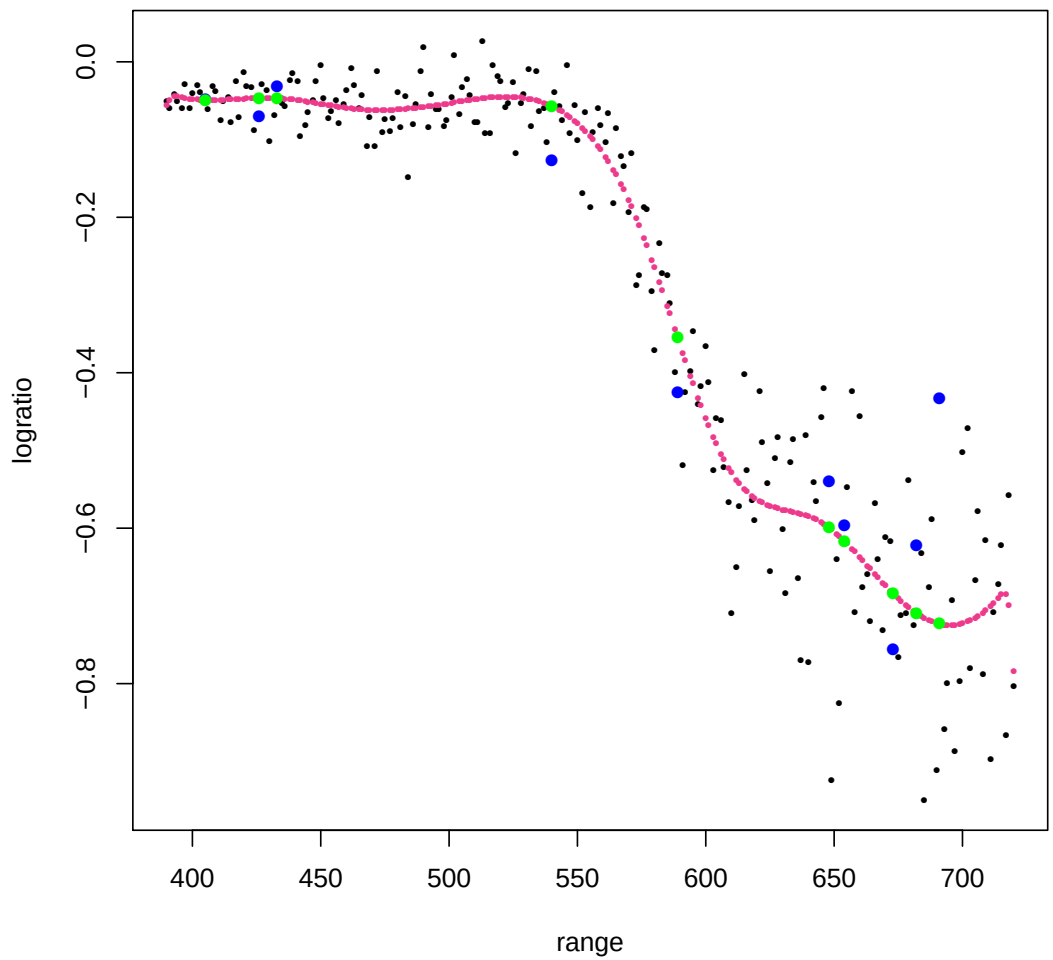
- Sélection de modèle en régression
  - Stepwise, forward, backward
- Échantillon test

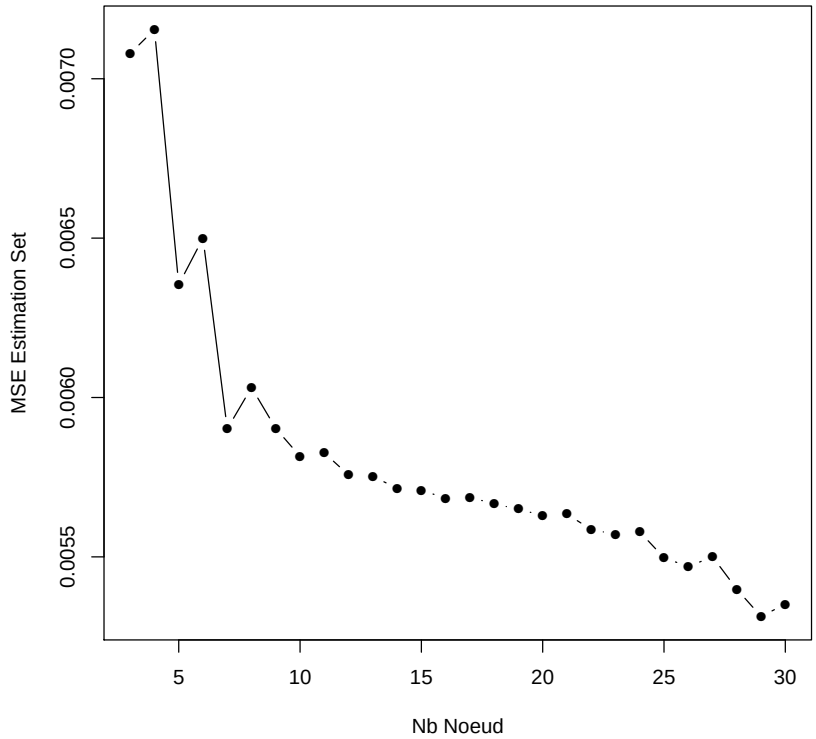
*Méthode de Pénalisation*

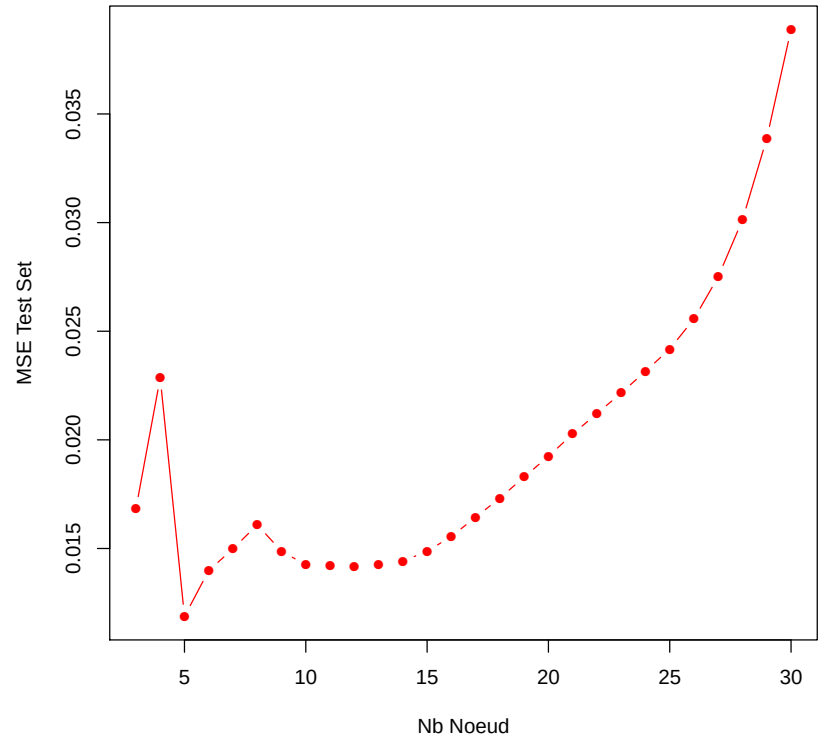
- AIC, BIC, Cp
- Pénalisation $L^1$: LASSO
- Pénalisation $L^2$: **Ridge regression**, **P-spline**


logratio
0.0
−0.2
−0.4
−0.6
−0.8
400
450
500
550
600
650
700
range

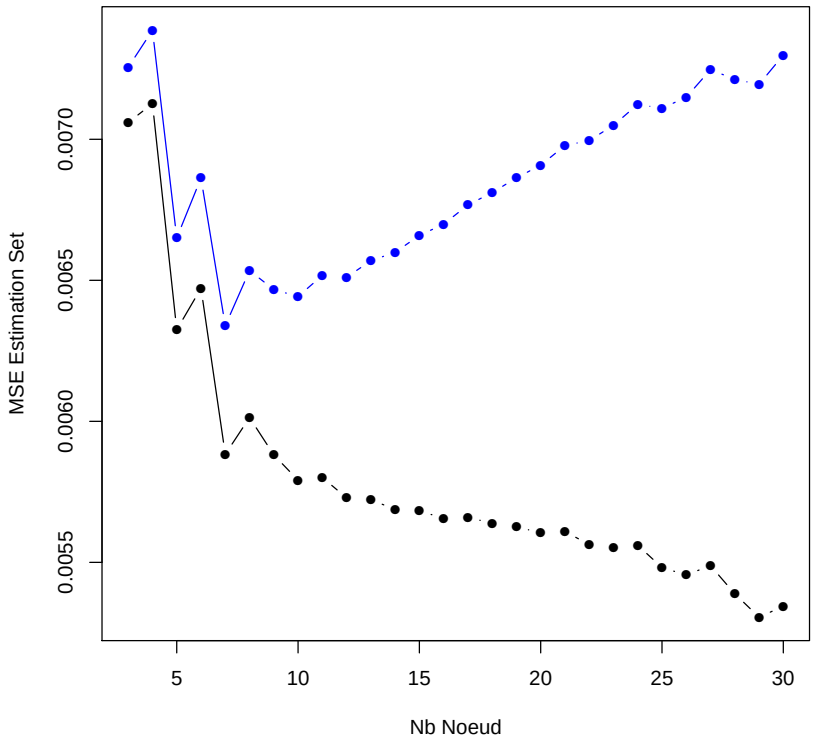

logratio
0.0
−0.2
−0.4
−0.6
−0.8
400
450
500
550
600
650
700
range

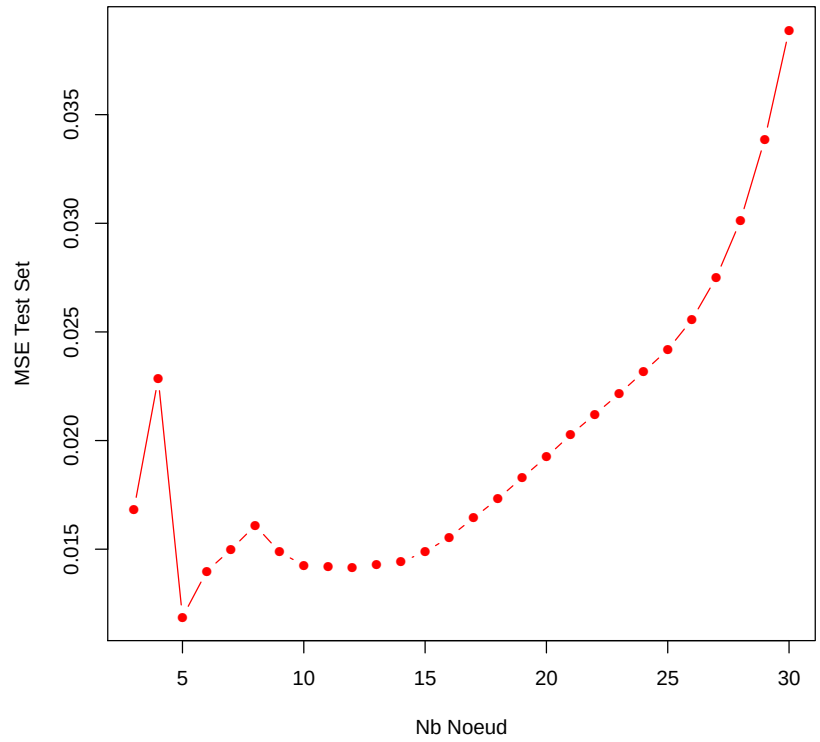

logratio
0.0
−0.2
−0.4
−0.6
−0.8
400
450
500
550
600
650
700
range




$\Rightarrow$ Pour éviter le **surapprentissage**, il faut ajouter une pénalité ou estimer l'erreur de prévision par validation croisée, échantillon test...


MSE Estimation Set
0.0055
0.0060
0.0065
0.0070
5
10
15
20
25
30
Nb Noeud


MSE Test Set
0.015
0.020
0.025
0.030
0.035
5
10
15
20
25
30
Nb Noeud